CA

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 15/11, 15/10, C07K 14/47, C12P 21/00, C12Q 1/68, C07K 16/18, G06F 17/30, 17/50

A2

(11) International Publication Number: **WO 99/53051**

(43) International Publication Date: 21 October 1999 (21.10.99)

(21) International Application Number: PCT/IB99/00712

(22) International Filing Date: 9 April 1999 (09.04.99)

(30) Priority Data:

09/057,719	9 April 1998 (09.04.98)	US
09/069,047	28 April 1998 (28.04.98)	US

(71) Applicant *(for all designated States except US)*: GENSET [FR/FR]; 24, rue Royale, F–75008 Paris (FR).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: DUMAS MILNE EDWARDS, Jean–Baptiste [FR/FR]; 8, rue Grégoire–de–Tours, F–75006 Paris (FR). DUCLERT, Aymeric [FR/FR]; 6 ter, rue Victorine, F–94100 Saint–Maur (FR). GIORDANO, Jean–Yves [FR/FR]; 12, rue Duhesme, F–75018 Paris (FR).

(74) Agents: MARTIN, Jean–Jacques et al.; Cabinet Regimbeau, 26, avenue Kléber, F–75116 Paris (FR).

(81) Designated States: AU, CA, JP, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: 5' ESTS AND ENCODED HUMAN PROTEINS

(57) Abstract

The sequences of 5' ESTs derived from mRNAs encoding secreted proteins are disclosed. The 5' ESTs may be to obtain cDNAs and genomic DNAs corresponding to the 5' ESTs. The 5' ESTs may also be used in diagnostic, forensic, gene therapy, and chromosome mapping procedures. Upstream regulatory sequences may also be otained using the 5' ESTs. The 5' ESTs may also be used to design expression vectors and secretion vectors.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

5' ESTS AND ENCODED HUMAN PROTEINS

Background of the Invention

The estimated 50,000-100,000 genes scattered along the human chromosomes offer tremendous promise for the understanding, diagnosis, and treatment of human diseases. In addition, probes capable of specifically hybridizing to loci distributed throughout the human genome find applications in the construction of high resolution chromosome maps and in the identification of individuals.

In the past, the characterization of even a single human gene was a painstaking process, requiring years of effort. Recent developments in the areas of cloning vectors, DNA sequencing, and computer technology have merged to greatly accelerate the rate at which human genes can be isolated, sequenced, mapped, and characterized.

Currently, two different approaches are being pursued for identifying and characterizing the genes distributed along the human genome. In one approach, large fragments of genomic DNA are isolated, cloned, and sequenced. Potential open reading frames in these genomic sequences are identified using bioinformatics software. However, this approach entails sequencing large stretches of human DNA which do not encode proteins in order to find the protein encoding sequences scattered throughout the genome. In addition to requiring extensive sequencing, the bioinformatics software may mischaracterize the genomic sequences obtained, *i.e.*, labeling non-coding DNA as coding DNA and vice versa.

An alternative approach takes a more direct route to identifying and characterizing human genes. In this approach, complementary DNAs (cDNAs) are synthesized from isolated messenger RNAs (mRNAs) which encode human proteins. Using this approach, sequencing is only performed on DNA which is derived from protein coding portions of the genome. Often, only short stretches of the cDNAs are sequenced to obtain sequences called expressed sequence tags (ESTs). The ESTs may then be used to isolate or purify extended cDNAs which include sequences adjacent to the EST sequences. The extended cDNAs may contain all of the sequence of the EST which was used to obtain them or only a portion of the sequence of the EST which was used to obtain them. In addition, the extended cDNAs may contain the full coding sequence of the gene from which the EST was derived or, alternatively, the extended cDNAs may include portions of the coding sequence of the gene from which the EST was derived. It will be appreciated that there may be several extended cDNAs which include the EST sequence as a result of alternate splicing or the activity of alternative promoters. Alternatively, ESTs having partially overlapping sequences may be identified and contigs comprising the consensus sequences of the overlapping ESTs may be identified.

In the past, these short EST sequences were often obtained from oligo-dT primed cDNA libraries. Accordingly, they mainly corresponded to the 3' untranslated region of the mRNA. In part, the prevalence of EST sequences derived from the 3' end of the mRNA is a result of the fact that typical

techniques for obtaining cDNAs, are not well suited for isolating cDNA sequences derived from the 5' ends of mRNAs (Adams *et al., Nature* 377:3-174, 1996, Hillier *et al., Genome Res.* 6:807-828, 1996).

In addition, in those reported instances where longer cDNA sequences have been obtained, the reported sequences typically correspond to coding sequences and do not include the full 5' untranslated region (5'UTR) of the mRNA from which the cDNA is derived. Indeed, 5'UTRs have been shown to affect either the stability or translation of mRNAs. Thus, regulation of gene expression may be achieved through the use of alternative 5'UTRs as shown, for instance, for the translation of the tissue inhibitor of metalloprotease mRNA in mitogenically activated cells (Waterhouse *et al, J Biol Chem.* **265**:5585-9. 1990). Furthermore, modification of 5'UTR through mutation, insertion or translocation events may even be implied in pathogenesis. For instance, the fragile X syndrome, the most common cause of inherited mental retardation, is partly due to an insertion of multiple CGG trinucleotides in the 5'UTR of the fragile X mRNA resulting in the inhibition of protein synthesis via ribosome stalling (Feng *et al, Science* **268**:731-4, 1995). An aberrant mutation in regions of the 5'UTR known to inhibit translation of the proto-oncogene *c-myc* was shown to result in upregulation of c-myc protein levels in cells derived from patients with multiple myelomas (Willis *et al, Curr Top Microbiol Immunol* **224**:269-76, 1997). In addition, the use of oligo-dT primed cDNA libraries does not allow the isolation of complete 5'UTRs since such incomplete sequences obtained by this process may not include the first exon of the mRNA, particularly in situations where the first exon is short. Furthermore, they may not include some exons, often short ones, which are located upstream of splicing sites. Thus, there is a need to obtain sequences derived from the 5' ends of mRNAs.

While many sequences derived from human chromosomes have practical applications, approaches based on the identification and characterization of those chromosomal sequences which encode a protein product are particularly relevant to diagnostic and therapeutic uses. In some instances, the sequences used in such therapeutic or diagnostic techniques may be sequences which encode proteins which are secreted from the cell in which they are synthesized. Those sequences encoding secreted proteins as well as the secreted proteins themselves, are particularly valuable as potential therapeutic agents. Such proteins are often involved in cell to cell communication and may be responsible for producing a clinically relevant response in their target cells. In fact, several secretory proteins, including tissue plasminogen activator, G-CSF, GM-CSF, erythropoietin, human growth hormone, insulin, interferon-α, interferon-β, interferon-γ, and interleukin-2, are currently in clinical use. These proteins are used to treat a wide range of conditions, including acute myocardial infarction, acute ischemic stroke, anemia, diabetes, growth hormone deficiency, hepatitis, kidney carcinoma, chemotherapy-induced neutropenia and multiple sclerosis. For these reasons, extended cDNAs encoding secreted proteins or portions thereof represent a valuable source of therapeutic agents. Thus, there is a need for the identification and characterization of secreted proteins and the nucleic acids encoding them.

In addition to being therapeutically useful themselves, secretory proteins include short peptides, called signal peptides, at their amino termini which direct their secretion. These signal peptides are

encoded by the signal sequences located at the 5' ends of the coding sequences of genes encoding secreted proteins. These signal peptides can be used to direct the extracellular secretion of any protein to which they are operably linked. In addition, portions of the signal peptides called membrane-translocating sequences, may also be used to direct the intracellular import of a peptide or protein of interest. This may prove beneficial in gene therapy strategies in which it is desired to deliver a particular gene product to cells other than the cells in which it is produced. Signal sequences encoding signal peptides also find application in simplifying protein purification techniques. In such applications, the extracellular secretion of the desired protein greatly facilitates purification by reducing the number of undesired proteins from which the desired protein must be selected. Thus, there exists a need to identify and characterize the 5' portions of the genes for secretory proteins which encode signal peptides.

Sequences coding for non-secreted proteins may also find application as therapeutics or diagnostics. In particular, such sequences may be used to determine whether an individual is likely to express a detectable phenotype, such as a disease, as a consequence of a mutation in the coding sequence of a protein. In instances where the individual is at risk of suffering from a disease or other undesirable phenotype as a result of a mutation in such a coding sequence, the undesirable phenotype may be corrected by introducing a normal coding sequence using gene therapy. Alternatively, if the undesirable phenotype results from overexpression of the protein encoded by the coding sequence, expression of the protein may be reduced using antisense or triple helix based strategies.

The secreted or non-secreted human polypeptides encoded by the coding sequences may also be used as therapeutics by administering them directly to an individual having a condition, such as a disease, resulting from a mutation in the sequence encoding the polypeptide. In such an instance, the condition can be cured or ameliorated by administering the polypeptide to the individual.

In addition, the secreted or non-secreted human polypeptides or portions thereof may be used to generate antibodies useful in determining the tissue type or species of origin of a biological sample. The antibodies may also be used to determine the cellular localization of the secreted or non-secreted human polypeptides or the cellular localization of polypeptides which have been fused to the human polypeptides. In addition, the antibodies may also be used in immunoaffinity chromatography techniques to isolate, purify, or enrich the human polypeptide or a target polypeptide which has been fused to the human polypeptide.

Public information on the number of human genes for which the promoters and upstream regulatory regions have been identified and characterized is quite limited. In part, this may be due to the difficulty of isolating such regulatory sequences. Upstream regulatory sequences such as transcription factor binding sites are typically too short to be utilized as probes for isolating promoters from human genomic libraries. Recently, some approaches have been developed to isolate human promoters. One of them consists of making a CpG island library (Cross *et al., Nature Genetics* 6: 236-244, 1994). The second consists of isolating human genomic DNA sequences containing SpeI binding sites by the use of SpeI binding protein. (Mortlock *et al., Genome Res.* 6:327-335, 1996). Both of these approaches have

their limits due to a lack of specificity and of comprehensiveness. Thus, there exists a need to identify and systematically characterize the 5' portions of the genes.

The present 5' ESTs may be used to efficiently identify and isolate 5'UTRs and upstream regulatory regions which control the location, developmental stage, rate, and quantity of protein synthesis, as well as the stability of the mRNA. Once identified and characterized, these regulatory regions may be utilized in gene therapy or protein purification schemes to obtain the desired amount and locations of protein synthesis or to inhibit, reduce, or prevent the synthesis of undesirable gene products.

In addition, ESTs containing the 5' ends of protein genes may include sequences useful as probes for chromosome mapping and the identification of individuals. Thus, there is a need to identify and characterize the sequences upstream of the 5' coding sequences of genes.

Summary of the Invention

The present invention relates to purified, isolated, or enriched 5' ESTs which include sequences derived from the authentic 5' ends of their corresponding mRNAs. The term "corresponding mRNA" refers to the mRNA which was the template for the cDNA synthesis which produced the 5' EST. These sequences will be referred to hereinafter as "5' ESTs." The present invention also includes purified, isolated or enriched nucleic acids comprising contigs assembled by determining a consensus sequences from a plurality of ESTs containing overlapping sequences. These contigs will be referred to herein as "consensus contigated 5'ESTs."

As used herein, the term "purified" does not require absolute purity; rather, it is intended as a relative definition. Individual 5' EST clones isolated from a cDNA library have been conventionally purified to electrophoretic homogeneity. The sequences obtained from these clones could not be obtained directly either from the library or from total human DNA. The cDNA clones are not naturally occurring as such, but rather are obtained via manipulation of a partially purified naturally occurring substance (messenger RNA). The conversion of mRNA into a cDNA library involves the creation of a synthetic substance (cDNA) and pure individual cDNA clones can be isolated from the synthetic library by clonal selection. Thus, creating a cDNA library from messenger RNA and subsequently isolating individual clones from that library results in an approximately 10^4-10^6 fold purification of the native message. Purification of starting material or natural material to at least one order of magnitude, preferably two or three orders, and more preferably four or five orders of magnitude is expressly contemplated.

As used herein, the term "isolated" requires that the material be removed from its original environment (e.g., the natural environment if it is naturally occurring). For example, a naturally-occurring polynucleotide present in a living animal is not isolated, but the same polynucleotide, separated from some or all of the coexisting materials in the natural system, is isolated.

As used herein, the term "recombinant" means that the 5' EST is adjacent to "backbone" nucleic acid to which it is not adjacent in its natural environment. Additionally, to be "enriched" the 5' ESTs will

Brief Description of the Drawings

Figure 1 is a summary of a procedure for obtaining cDNAs which have been selected to include the 5′ ends of the mRNAs from which they derived. In the first step (1), the cap of intact mRNAs is oxidized to be chemically ligated to an oligonucleotide tag. In the second step (2), a reverse transcription is performed using random primers to generate a first cDNA strand. In the third step (3), mRNAs are eliminated and the second strand synthesis is carried out using a primer contained in the oligonucleotide tag.

Figure 2 is an analysis of the 43 amino terminal amino acids of all human SwissProt proteins to determine the frequency of false positives and false negatives using the techniques for signal peptide identification described herein.

Figure 3 summarizes a general method used to clone and sequence extended cDNAs containing sequences adjacent to 5'ESTs.

Figure 4 provides a schematic description of the promoters isolated and the way they are assembled with the corresponding 5' tags.

Figure 5 describes the transcription factor binding sites present in each of the promoters of Figure 4.

Figure 6 is a block diagram of an exemplary computer system.

Figure 7 is a flow diagram illustrating one embodiment of a process 200 for comparing a new nucleotide or protein sequence with a database of sequences in order to determine the homology levels between the new sequence and the sequences in the database.

Figure 8 is a flow diagram illustrating one embodiment of a process 250 in a computer for determining whether two sequences are homologous.

Figure 9 is a flow diagram illustrating one embodiment of an identifier process 300 for detecting the presence of a feature in a sequence.

Figure 10 is a table with all of the parameters that can be used for each step of extended cDNA analysis.

Detailed Description of the Preferred Embodiment

I. Obtaining 5'ESTs from cDNA libraries including the 5'Ends of their Corresponding mRNAs

The 5' ESTs of the present invention were obtained from cDNA libraries including cDNAs which include the 5'end of their corresponding mRNAs. The general method used to obtain such cDNA libraries is described in Examples 1 to 5.

EXAMPLE 1

Preparation of mRNA

Total human RNAs or polyA$^+$ RNAs derived from 29 different tissues were respectively purchased from LABIMO and CLONTECH and used to generate 44 cDNA libraries as described below.

CLAIMS

1. A purified nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and sequences complementary to the sequences of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622.

2. A purified nucleic acid comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and sequences complementary to the sequences of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622.

3. A purified or isolated polypeptide comprising a sequence selected from the group consisting of the sequences of SEQ ID NOs. 812-1599.

4. A method of making a cDNA comprising the steps of:

a) contacting a collection of mRNA molecules from human cells with a primer comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of the sequences complementary to SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622;

b) hybridizing said primer to an mRNA in said collection that encodes said protein;

c) reverse transcribing said hybridized primer to make a first cDNA strand from said mRNA;

d) making a second cDNA strand complementary to said first cDNA strand; and

e) isolating the resulting cDNA comprising said first cDNA strand and said second cDNA strand.

5. A method of making a cDNA comprising the steps of:

a) obtaining a cDNA comprising a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622;

b) contacting said cDNA with a detectable probe comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and the sequences complementary to SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 under conditions which permit said probe to hybridize to said cDNA;

c) identifying a cDNA which hybridizes to said detectable probe; and

d) isolating said cDNA which hybridizes to said probe.

6. A method of making a cDNA comprising the steps of:

a) contacting a collection of mRNA molecules from human cells with a first primer capable of hybridizing to the polyA tail of said mRNA;

b) hybridizing said first primer to said polyA tail;

c) reverse transcribing said mRNA to make a first cDNA strand;

d) making a second cDNA strand complementary to said first cDNA strand using at least one primer comprising at least 15 consecutive nucleotides of a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622; and

e) isolating the resulting cDNA comprising said first cDNA strand and said second cDNA strand.

7. A method of making a polypeptide comprising the steps of:

a) obtaining a cDNA which encodes a polypeptide encoded by a nucleic acid comprising a sequence selected from the group consisting of SEQ ID NOs. 24-811 or a cDNA which encodes a polypeptide comprising at least 10 consecutive amino acids of a polypeptide encoded by a sequence selected from the group consisting of SEQ ID NOs. 24-811;

b) inserting said cDNA in an expression vector such that said cDNA is operably linked to a promoter;

c) introducing said expression vector into a host cell whereby said host cell produces the protein encoded by said cDNA; and

d) isolating said protein.

8. In an array of discrete ESTs or fragments thereof of at least 15 nucleotides in length, the improvement comprising inclusion in said array of at least one sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622, the sequences complementary to the sequences of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and fragments comprising at least 15 consecutive nucleotides of said sequence.

9. The array of Claim 8 including therein at least five sequences selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622, the sequences complementary to the sequences of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and fragments comprising at least 15 consecutive nucleotides of said sequences.

10. An enriched population of recombinant nucleic acids, said recombinant nucleic acids comprising an insert nucleic acid and a backbone nucleic acid, wherein at least 5% of said insert nucleic acids in said population comprise a sequence selected from the group consisting of SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622, the sequences complementary to SEQ ID NOs. 24-811 and SEQ ID NOs. 1600-1622 and fragments comprising at least 15 consecutive nucleotides of said sequences.

11. An antibody composition capable of selectively binding to an epitope-containing fragment of a polypeptide comprising a contiguous span of at least 8 amino acids of any of SEQ ID NOs. 812-1599, wherein said antibody is polyclonal or monoclonal.

12. A computer readable medium having stored thereon a sequence selected from the group consisting of a nucleic acid code of SEQ ID NOs. 24-811 and 1600-1622 and a polypeptide code of SEQ ID NOs. 812-1599.

13. A computer system comprising a processor and a data storage device wherein said data storage device has stored thereon a sequence selected from the group consisting of a nucleic acid code of SEQID NOs. 24-811 and 1600-1622 and a polypeptide code of SEQ ID NOs. 812-1599.

14. The computer system of Claim 13 further comprising a sequence comparer and a data storage device having reference sequences stored thereon.

15. The computer system of Claim 14 wherein said sequence comparer comprises a computer program which indicates polymorphisms.

16. The computer system of Claim 13 further comprising an identifier which identifies features in said sequence.

17. A method for comparing a first sequence to a reference sequence wherein said first sequence is selected from the group consisting of a nucleic acid code of SEQID NOs. 24-811 and 1600-1622 and a polypeptide code of SEQ ID NOs. 812-1599 comprising the steps of:

a) reading said first sequence and said reference sequence through use of a computer program which compares sequences; and

b) determining differences between said first sequence and said reference sequence with said computer program.

18. The method of Claim 17, wherein said step of determining differences between the first sequence and the reference sequence comprises identifying polymorphisms.

19. A method for identifying a feature in a sequence selected from the group consisting of a nucleic acid code of SEQID NOs. 24-811 and 1600-1622 and a polypeptide code of SEQ ID NOs. 812-1599 comprising the steps of:

a) reading said sequence through the use of a computer program which identifies features in sequences; and

b) identifying features in said sequence with said computer program.

20. A vector comprising a nucleic acid according to either Claims 1 or 2.

21. A host cell containing a nucleic acid of Claim 20.

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12N 15/11, 15/10, C07K 14/47, C12P 21/00, C12Q 1/68, C07K 16/18, G06F 17/30, 17/50

A3

(11) International Publication Number: **WO 99/53051**

(43) International Publication Date: 21 October 1999 (21.10.99)

(21) International Application Number: PCT/IB99/00712

(22) International Filing Date: 9 April 1999 (09.04.99)

(30) Priority Data:

09/057,719	9 April 1998 (09.04.98)	US
09/069,047	28 April 1998 (28.04.98)	US

(71) Applicant *(for all designated States except US)*: GENSET [FR/FR]; 24, rue Royale, F–75008 Paris (FR).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: DUMAS MILNE EDWARDS, Jean–Baptiste [FR/FR]; 8, rue Grégoire–de–Tours, F–75006 Paris (FR). DUCLERT, Aymeric [FR/FR]; 6 ter, rue Victorine, F–94100 Saint–Maur (FR). GIORDANO, Jean–Yves [FR/FR]; 12, rue Duhesme, F–75018 Paris (FR).

(74) Agents: MARTIN, Jean–Jacques et al.; Cabinet Regimbeau, 26, avenue Kléber, F–75116 Paris (FR).

(81) Designated States: AU, CA, JP, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

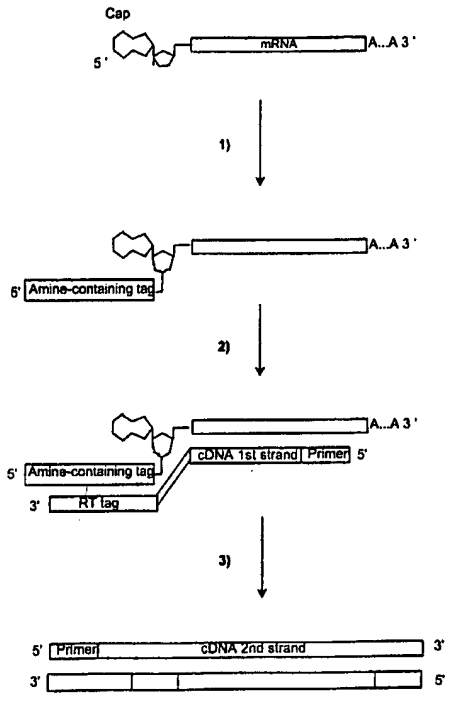
Published
With international search report.

(88) Date of publication of the international search report:
6 April 2000 (06.04.00)

(54) Title: 5' ESTS AND ENCODED HUMAN PROTEINS

(57) Abstract

The sequences of 5' ESTs derived from mRNAs encoding secreted proteins are disclosed. The 5' ESTs may be to obtain cDNAs and genomic DNAs corresponding to the 5' ESTs. The 5' ESTs may also be used in diagnostic, forensic, gene therapy, and chromosome mapping procedures. Upstream regulatory sequences may also be otained using the 5' ESTs. The 5' ESTs may also be used to design expression vectors and secretion vectors.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

INTERNATIONAL SEARCH REPORT

Interna ial Application No
PCT/IB 99/00712

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 C12N15/11 C12N15/10 C07K14/47 C12P21/00 C12Q1/68
C07K16/18 G06F17/30 G06F17/50

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 C12N C07K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	BRENNER ET AL.: "Homo sapiens Xq28 genomic DNA in the region of the L1CAM locus ...." EMBL SEQUENCE DATABASE, 9 May 1996 (1996-05-09), XP002121588 HEIDELBERG DE	1,2
Y	Ac U52112 the whole document & BRENNER ET AL.: "Genomic organization of two novel genes on human Xq28: compact head to head arrangement of IDH gamma and TRAP delta is conserved in rat and mouse" GENOMICS, vol. 44, no. 1, 1997, pages 8-14, --- -/--	4-21

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
4 November 1999	28. JAN. 2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	CEDER O.

2

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SAKAI ET AL.: "Protein kinase C substrate, 80 kD protein, heavy chain (PKCSH)" SWISSPROT SEQUENCE DATA BASE, 1 January 1990 (1990-01-01), XP002121589 HEIDELBERG DE Ac P14314 the whole document -& SAKAI ET AL.: "Human 80K-H protein (kinase C substrate) mRNA, complete compound" EMBL SEQUENCE DATABASE, 1 February 1990 (1990-02-01), XP002121590 HEIDELBERG DE Ac J03075 the whole document & SAKAI ET AL.: "Isolation of cDNAs encoding a substrate for protein kinase C: nucleotide sequence and chromosomal mapping of the gene for a human 80K protein" GENOMICS, vol. 5, 1989, pages 309-315, ---	3
Y	WO 96 34981 A (GENSET ;MERENKOVA IRENA NICOLAEVNA (FR); DUMAS MILNE EDWARDS JEAN) 7 November 1996 (1996-11-07) cited in the application page 13, line 24 -page 14, line 14; claim 26 ---	4
Y	GREENWOOD M T ET AL: "Cloning of the gene encoding human somatostatin receptor 2: sequence analysis of the 5@?-flanking promoter region" GENE, vol. 159, no. 2, 4 July 1995 (1995-07-04), page 291-292 XP004042228 ISSN: 0378-1119 abstract ---	5
Y	KATO S ET AL: "Construction of a human full-length cDNA bank" GENE, vol. 150, 1 January 1994 (1994-01-01), pages 243-250, XP002081364 ISSN: 0378-1119 cited in the application abstract page 245, left-hand column --- -/--	6,10

INTERNATIONAL SEARCH REPORT

Intern. nal Application No
PCT/IB 99/00712

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y A	WO 97 38003 A (HUMAN GENOME SCIENCES INC ;LI HAODONG (US); WEI YING FEI (US)) 16 October 1997 (1997-10-16) seq Id No 2 claims 10-12	7,11,20, 21 3
Y	LOCKHART D J ET AL: "EXPRESSION MONITORING BY HYBRIDIZATION TO HIGH-DENSITY OLIGONUCLEOTIDE ARRAYS" BIO/TECHNOLOGY, vol. 14, no. 13, 1 December 1996 (1996-12-01), pages 1675-1680, XP002022521 ISSN: 0733-222X abstract	8,9
Y	WO 98 07830 A (INST GENOMIC RESEARCH ;UNIV PENNSYLVANIA (US); UNIV JOHNS HOPKINS) 26 February 1998 (1998-02-26) page 3, line 4 - line 28 page 31, line 6 -page 35, line 16	7,11-21
X	MUZNY ET AL.: "Homo sapiens, working draft sequence, 97 unordered pieces" EMBL SEQUENCE DATABASE, 3 February 1998 (1998-02-03), XP002121591 HEIDELBERG DE Ac AC004085 the whole document	1,2
X	ADAMS ET AL.: "EST177394 Jurkat T-cells VI homo sapiens cDNA 5' end similar to protein kinase C substrate 80K-H" EMBL SEQUENCE DATABASE, 18 April 1997 (1997-04-18), XP002121592 HEIDELBERG DE Ac AA306438 the whole document -& ADAMS ET AL.: "Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequences" NATURE, vol. 377, 1995, pages 3-174, XP002069461	3
A	"zr94d07.r1 NCI_CGAP_GCB1 Homo sapiens cDNA clone IMAGE:683341 5' EST" EMBL SEQUENCE DATABASE, 5 February 1997 (1997-02-05), XP002121593 HEIDELBERG DE Ac AA215334 the whole document	1,2

-/--

2

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Intern ıal Application No
PCT/IB 99/00712

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	ADAMS M D ET AL: "RAPID CDNA SEQUENCING (EXPRESSED SEQUENCE TAGS) FROM A DIRECTIONALLY CLONED HUMAN INFANT BRAIN CDNA LIBRARY" NATURE GENETICS, vol. 4, no. 4, 1 August 1993 (1993-08-01), pages 373-380, STANDARD, XP002064427 ISSN: 1061-4036 ---	
A	ADAMS M D ET AL: "3,400 NEW EXPRESSED SEQUENCE TAGS IDENTIFY DIVERSITY OF TRANSCRIPTS IN HUMAN BRAIN" NATURE GENETICS, vol. 4, no. 3, 1 July 1993 (1993-07-01), pages 256-267, XP000645060 ISSN: 1061-4036 ---	
A	TASHIRO K ET AL: "SIGNAL SEQUENCE TRAP: A CLONING STRATEGY FOR SECRETED PROTEINS AND TYPE I MEMBRANE PROTEINS" SCIENCE, vol. 261, 30 July 1993 (1993-07-30), pages 600-603, XP000673204 ISSN: 0036-8075 ---	
A	CARNINCI P ET AL: "High-efficiency full-length cDNA cloning by biotinylated CAP trapper" GENOMICS, vol. 37, no. 3, 1 November 1996 (1996-11-01), pages 327-336, XP002081729 ISSN: 0888-7543 -----	

2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IB 99/00712

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. [X] Claims Nos.: -
because they relate to subject matter not required to be searched by this Authority, namely:

Rule 39.1(v) PCT - Presentation of information
Although claim 12 could be considered as a mere presentation of information, Rule 39.1(v) PCT, the search has been carried out as far as possible in our systematic documentation.

2. [ ] Claims Nos.:
because they relate to parts of the International Application that do not comply with the prescribed requirements to such an extent that no meaningful International Search can be carried out, specifically:

3. [ ] Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. [ ] As all required additional search fees were timely paid by the applicant, this International Search Report covers all searchable claims.

2. [ ] As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. [ ] As only some of the required additional search fees were timely paid by the applicant, this International Search Report covers only those claims for which fees were paid, specifically claims Nos.:

4. [X] No required additional search fees were timely paid by the applicant. Consequently, this International Search Report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Invention 1: 1-21 partially

Remark on Protest

[ ] The additional search fees were accompanied by the applicant's protest.

[ ] No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Continuation of Box I.1

Although claim 12 could be considered as a mere presentation of information, Rule 39.1(v) PCT, the search has been carried out as far as possible in our systematic documentation.

Continuation of Box I.1

Rule 39.1(v) PCT - Presentation of information

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

1. Claims: Invention 1: 1-21 all partially

Nucleic acid comprising a sequence as in Seq.ID.No. 24, complementary sequence and fragments thereof. Polypeptide, Seq.Id.No. 812, encoded by said nucleotide sequence. Vector comprising Seq.Id.No. 24 and host cell comprising the vector. Methods of making cDNA and polypeptide utilising Seq.Id.No. 24. Array of ESTs comprising Seq.Id.No. 24, or a fragment thereof. An antibody binding to an epitop of the polypeptide of Seq.Id.No. 812. A computer readable medium and a computer system storing and/or utilising the sequence of Seq.Id.No. 24 or 812.

2. Claims: Invention 2-811 : 1-21 all partially

Idem as subject 1 but limited to each of the DNA sequences as in Seq.Id.No. 25-811 and 1600-1622, and corresponding polypeptides when applicible, where invention 2 is limited to Seq.Id.No. 25 and 813, invention 3 is limited to Seq.Id.No. 26 and 814, ..., invention 788 is limited to Seq.Id.No. 811 and 1599, invention 789 is limited to Seq.Id.No. 1600, invention 790 is limited to Seq.Id.No. 1601, ..., invention 811 is limited to Seq.Id.No. 1622.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/IB 99/00712

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9634981 A	07-11-1996	FR 2733765 A	08-11-1996
		FR 2733762 A	08-11-1996
		AU 5982996 A	21-11-1996
		CA 2220045 A	07-11-1996
		EP 0824598 A	25-02-1996
		JP 11510364 T	14-09-1999
WO 9738003 A	16-10-1997	AU 5389096 A	29-10-1997
		US 5945303 A	31-08-1999
WO 9807830 A	26-02-1998	NONE	